A Protein Family Classification Method for Analysis of Large DNA Sequences

Steven Henikoff[1] and Jorja G. Henikoff

[1]Howard Hughes Medical Institute Basic Sciences Division
Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Abstract

A method is described for identification and classification of proteins encoded in large DNA sequences. Previously, an automated system was introduced for the general detection of amino acid sequence motifs within diverse protein families. The system generated a database consisting of aligned sequence segments (blocks) that correspond to the most highly conserved regions of proteins. This database of blocks can be searched using protein queries for sensitive detection of homology based on the detection of both local and global similarities. Here we show that this database searching approach can also be used to detect distant relatives encoded in very large DNA sequences. The approach is illustrated by the detection of known and new relationships in the 315 kilobase (kb) sequence of yeast chromosome III.

1: Introduction

Automation and new approaches applied to DNA sequence acquisition methodology have led to an increase in the size and number of large sequencing projects in recent years, a trend that is likely to continue. The motivation for these large-scale sequencing projects is typically to discover and analyze genes, a process that involves detection of protein homology, often the only or the most important clue to the function of a gene. The likelihood that useful homology will be detected for a new sequence increases as more and more genes of known function are sequenced, and their sequences placed in public databanks. However, this same increase in size of the databanks leads to higher background in searches, making distant relationships more difficult to detect with confidence. When this problem is encountered in large-scale projects, it is especially challenging because contextual information about any particular segment of DNA is lacking. This is in contrast to the situation for an individual investigator interested in a single gene, who usually has biological insights that allow informed judgments to be made. For this reason, simplified automated approaches to sequence analysis are especially important for the interpretation of large-scale sequence data [1], [2], [3], [4], [5].

Here we describe one such automated approach, and apply it to the largest available contiguous sequence in the current databanks, *Saccharomyces cerevisiae* chromosome III. We present results comparable to those obtained using labor-intensive manual approaches carried out by recognized experts in sequence analysis [6], [7], [8] using standard similarity searching tools [9], [10], [11], [12], and even show the detection of homologies that were not reported by them.

2: Methods

2.1: A database of blocks

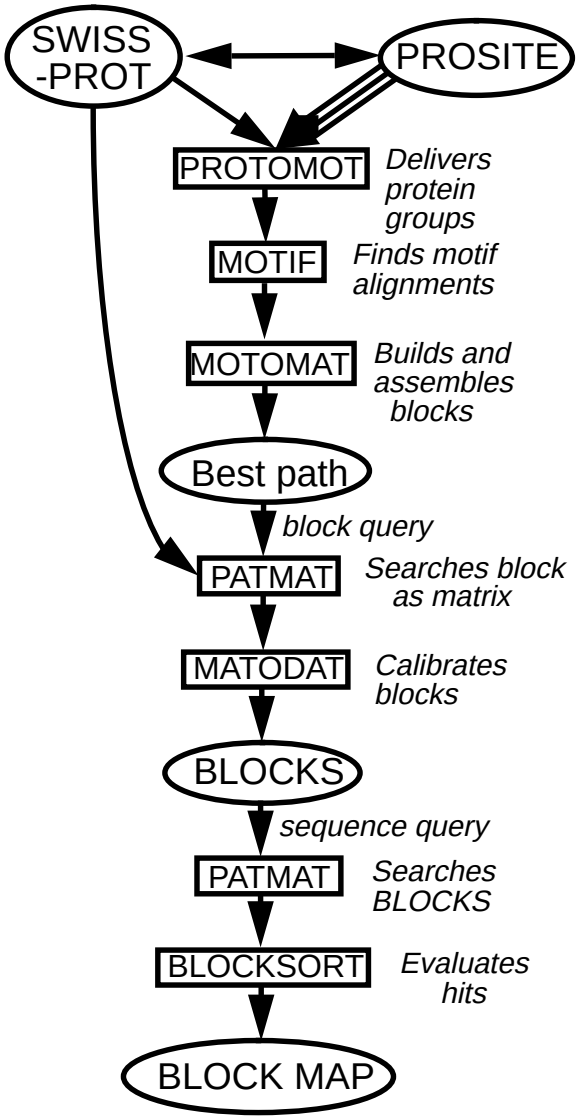

Fig. 1. Outline of the PROTOMAT system [2] applied to the PROSITE catalog keyed to the SWISS-PROT databank. The different modules are indicated, resulting in a set of blocks representing families in PROSITE, calibrated for searching using PATMAT.

Blocks consist of multiply aligned sequence segments without gaps. In previous work, we introduced the PROTOMAT system for making blocks from groups of related proteins [2]. The system consists of modules that can be executed singly or in combination, either manually using user-specified parameters or automatically using parameters determined by the system (Fig. 1). PROTOMAT is applied to successive groups, resulting in a database of blocks. For the blocks databases used in the current and previous studies (BLOCKS), groups consist of collections of proteins that share sequence similarity (and usually function) that are listed in the PROSITE catalog [13]. An entry in the PROSITE catalog contains the SWISS-PROT [14] IDs for members of a documented protein group.

First, the sequence extraction module of PROTOMAT (PROTOMOT) reads a PROSITE entry from the PROSITE.DAT file and finds the individual full-length sequences in the SWISS-PROT databank. Using these sequences, PROTOMOT then executes a modified version of Smith's

MOTIF program for finding motif alignments in an automatic mode, where all parameters are determined by the program [15]. MOTIF searches for any set of 3 amino acids separated by 2 fixed distances occurring in a large fraction of the protein sequences in the group, where all 20 amino acids and all possible distances up to 17 are considered. Each resulting motif forms the basis for an alignment, against which each of the remaining sequences is aligned. These block alignments are then refined by the MOTOMAT module which merges motifs if they are aligned identically in all of the sequences and extends alignments out in both directions until similarity declines. Segments are then clustered such that within a cluster, each segment is ≥80% identical to at least one other segment.

MOTOMAT then uses graph theory techniques to assemble "paths" of blocks, reporting a "best path". All blocks in a path must be in the same order without overlapping for all of the sequences represented, which might not include all sequences in the PROSITE group. Each block is written to a separate file in a format that resembles a PROSITE entry. For the sequences represented, the minimum and maximum distances from the preceding block are included in each block file.

Calibration of individual blocks and their concatenation into a single file is performed by the MATODAT module. First, the PATMAT searching program uses each block as a query of the SWISS-PROT database by converting the block to a position-specific scoring matrix [16] and scoring all possible alignments of the block and all sequences in the database [17]. MATODAT analyzes the rank-ordered search results to separate the scores of (true positive) sequences that were used to construct the block from the scores of other sequences (considered true negatives). A "lower calibration score" corresponding to the 99.5th percentile of true negative scores is used to normalize each block. The median of true positive scores provides an "upper calibration score"; dividing this by the lower calibration score and multiplying by 1000 provides a measure that is referred to as the "strength" of the block.

2.2: Searching a database of blocks

The Blocks Searcher system [18] consists of the successive execution of two programs. First, PATMAT converts each block to a position-specific scoring matrix and scores all possible alignments of the (translated) DNA or protein query sequence and the Blocks Database [17]. This procedure is the reverse of a sequence database search used for block calibration. For each position in the query sequence scores are assigned on a scale of 0-100%, reflecting the frequency of the residue at that position in the block, divided by the expected occurrence of the residue in proteins in general. For protein queries, amino acid frequencies in SWISS-PROT are used for weighting, whereas for translated DNA queries, which consist primarily of non-coding and out-of-frame sequence, codon frequencies are used. Stop codons are given zero score. The sum of scores for all aligned positions is divided by the lower calibration score and multiplied by 1000 to yield a normalized "PATMAT score". BLOCKSORT analyses the results of a PATMAT search by examining alignments of the query with multiple blocks representing a group.

A "hit" reported by the Blocks Searcher consists of at least one block with a PATMAT score >1000 from a protein group represented in the Blocks Database. The highest ranking block in a hit is called the "anchor" block and any other blocks in the hit are called "supporting" blocks. A supporting block must align with the query sequence in the correct order and within reasonable distance of the anchor block and higher ranking supporting blocks, as previously described [18].

The number of block alignments saved and reported by the Blocks Searcher depends on the length of the query sequence. The average protein is about 340 amino acids (aa), so that on the average there could be one protein-coding gene per 1000 bases (1kb) in gene-dense sequence. Accordingly, the Blocks Searcher reports up to as many hits as there are kb in the query sequence, with a minimum of 10. Since each group in the Blocks Database is represented by an average of 4

Table 1. Occurrence of hits for 7082 shuffled protein queries

log(E)	1/100 searches	1/1000 searches	1/7000 searches
-5	NA	NA	1040 (10.81)
-4	NA	1125 (55.72)	1307 (98.51)
-3	1054 (16.50)	1179 (80.24)	1312 (98.66)
-2	1173 (65.20)	1257 (95.78)	1422 (99.82)
-1	1203 (88.11)	1337 (99.10)	1403 (99.71)
0	1325 (98.89)	1480 (99.91)	1617 (100.00)

Values shown are the highest (1/7000 searches), 7th highest (1/1000 searches) and 70th highest (1/100 searches) PATMAT anchor block scores for hits with each value of log(E). The associated percentile value of all anchor block scores is in parentheses. NA: no hits were counted.

blocks, the Blocks Searcher saves 4 block alignments for each hit it reports. For a query the size of yeast chromosome III, with 315,357 bp, 4 x 316 = 1264 block alignments are saved and up to 316 hits are reported. For tests of shuffled sequence queries reported in this study, 5000 block alignments were saved by PATMAT and up to 1000 hits were reported by BLOCKSORT, allowing a more thorough examination of potential false positive alignments and hits.

A measure of local similarity is provided by a percentile score, which is obtained by comparing the PATMAT score of each anchor block to the distribution of PATMAT scores from 7082 searches of the Blocks Database using shuffled SWISS-PROT sequences [18]. For single block hits, this is the only measure of sequence similarity provided. However, for multiple block hits, an expectant value, E, is used to estimate the degree of global similarity. For example, a value of $E=10^{-3}$ is expected to occur by chance once for every 1000 searches of the database using a protein query of average length.

The expectant value E is computed as the product of probabilities for each supporting block, and each supporting block probability consists of the product of two parts. The first part estimates the probability that a supporting block achieves its rank among all the block alignments done for the search using a simple sampling without replacement model. The second part estimates the probability that a supporting block will lie within a reasonable distance of the anchor block on the query sequence using statistics from the sequences in the blocks for the protein group.

Hits from searches of shuffled protein sequences against the Blocks Database were used in our previous study to evaluate the expectant value for assessing global similarity for multiple block hits [18]. We used 7082 shuffled proteins present in SWISS-PROT 24, maintaining the average size and composition of proteins in general. For the resulting 43,783 hits, we found that the observed frequency of expectant values was very close to the magnitude of the expectant values.

An anchor block score and an expectant value for supporting blocks together provide independent evidence that can be used to evaluate a hit (Table 1), because the anchor block score is not used to calculate E. For example, a hit with anchor block score of 1312 (≥ the 98th percentile) and expectant value of 10^{-3} is expected to occur at least once by chance in 7000 searches using an average length protein, but is not expected to occur in 1000 searches.

2.3: Implementation

The Blocks Searcher has been implemented as an electronic mail server [19]. Detailed instructions with illustrative examples can be obtained by sending the message "help" in the subject line to blocks@howard.fhcrc.org. The Blocks Database is updated semi-annually following each significant update of Prosite. The current Blocks Database (BLOCKS v. 6.0) contains 2,302 calibrated blocks representing 619 groups. The PATMAT and BLOCKSORT programs are written in standard C for UNIX workstations and are available by anonymous ftp from the repository of the National Center for Biotechnology Information, ncbi.nlm.nih.gov, in the blocks subdirectory.

Further information can be obtained by sending a request to henikoff@howard.fhcrc.org.

3: Results

3.1: Extension of the Blocks Database searching system to large DNA sequences

In our previous study, we described the Blocks Searcher system for detecting both local and global homologies within protein sequences [18]. Here we are interested in homologies within very large DNA sequences that might contain multiple protein-coding genes. DNA queries are translated in all 6 frames and each translation is searched against the Blocks Database. Multiple block hits must be on the same strand, but not necessarily in the same frame. Evaluation is a problem because the Blocks Searcher presents results in terms of the expectation of a hit occurring by chance in the context of a single protein, not of a sequence that might encode hundreds of proteins. For example, the size of a protein equivalent to the 6-frame translation of yeast chromosome III is about 1850 times larger than a typical protein. This problem is illustrated in Fig. 2.

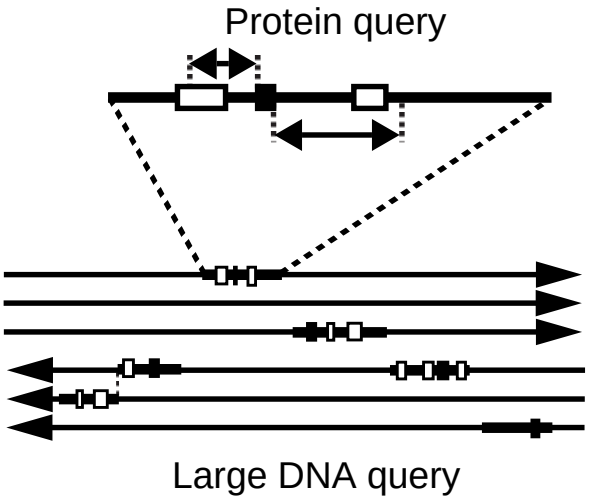

Fig. 2. Illustration of the conceptual problem addressed in this paper. Search results are understandable when interpreted in terms of what is found using protein query sequences of typical size (thick line). So, an anchor block (solid box) achieves a percentile score and supporting blocks (open boxes) achieve E-values based on rank and allowable distance between blocks (2-headed arrows) that are stated in terms of how frequently scores this high (or low) occur by chance in a typical search. But using a large DNA sequence translated in all 6 frames (arrows below) as query presents a problem, because there might be hundreds of proteins embedded within it, and the chance probability of obtaining any given score by chance is greatly increased in such a search. Still, searching in this way might be advantageous because it provides a single list of results, and because it makes no assumptions concerning ORFs, some of which might even be frameshifted (as illustrated at lower left). We therefore wish to evaluate these results as for single protein queries. This should be allowable if it can be shown that the E-values reported are realistic.

To test whether Blocks Searcher results are valid for large DNA queries, we shuffled *S. cerevisiae* chromosome III to provide large true negative queries (315,357 base pairs) for searching the Blocks Database. Shuffling was performed by randomly permuting individual bases, maintaining

Table 2. Relationship between expectant value (E) and observed hit frequency

Expectant value interval	Observed frequency and probability of hits within E interval			
	a) 7082 Shuffled Protein Queries		b) 70 Shuffled DNA Queries	
	Frequency	Observed P	Frequency	Observed P
$E \leq 3.5 \times 10^{-7}$	0	0	1	1.2×10^{-5}
$3.5 \times 10^{-7} < E \leq 3.5 \times 10^{-6}$	0	0	0	0
$3.5 \times 10^{-6} < E \leq 3.5 \times 10^{-5}$	1	2.3×10^{-5}	1	1.2×10^{-5}
$3.5 \times 10^{-5} < E \leq 3.5 \times 10^{-4}$	15	3.4×10^{-4}	26	3.0×10^{-4}
$3.5 \times 10^{-4} < E \leq 3.5 \times 10^{-3}$	124	2.8×10^{-3}	259	3.0×10^{-3}
$3.5 \times 10^{-3} < E \leq 3.5 \times 10^{-2}$	625	1.4×10^{-2}	60	6.9×10^{-4}
$3.5 \times 10^{-2} < E \leq 3.5 \times 10^{-1}$	3000	6.9×10^{-2}	0	0
$3.5 \times 10^{-1} < E$	40018	9.1×10^{-1}	86313	0.996
Total hits	43783		86600	

the overall base composition of the starting sequence. The gross alterations in coding regions that result from such shuffling should have only a slight effect on the analysis, because even for unshuffled sequences, 6-frame translation leads to about an order of magnitude more non-coding than coding sequence. Seventy different shuffles were performed, providing queries of composite length equivalent to about 130,000 different proteins of average size (70 shuffles x 315,357 bases ÷ 3 bases per aa x 6 frames ÷ 340 aa per protein = 129,853). This means that our searches involved 129,853 ÷ 7082 = 18 times the number of comparisons as in our previous experiments.

The results of the 70 searches using large shuffled DNA queries are compared with those using 7082 shuffled protein queries in Table 2. For the protein queries, an average of 340 amino acids x 2302 blocks = 782,680 block alignments were compared by PATMAT, and the top 400 (0.05%) were saved for analysis by BLOCKSORT, which found an average of about 6 hits per search. For the DNA queries, 315,357 bases x 2 strands x 2302 blocks = 1,451,903,628 block alignments were compared by PATMAT, and the top 5000 (0.0003%) were saved for analysis. BLOCKSORT was limited to reporting a maximum of 1000 hits, and found 1000 for every search. The minimum PATMAT score for an anchor block was 1021, well above 1001, the minimum anchor block score for a hit. This means that not enough PATMAT alignments and BLOCKSORT hits were saved to capture all possible hits, as was done for the protein queries, a reflection of the 1850-fold larger number of comparisons made when doing 6-frame translation searches of such a large sequence. In order to estimate the observed probability of the expectant (E) values for the DNA queries, we first estimated the total number of possible hits (2 strands x 619 families x 70 searches = 86,660 possible hits). The Blocks Searcher reported 70,000, or 81%, of these possible hits. Therefore, there should be nearly as many expectant values reported for the true negatives in the DNA searches as in the protein searches. Table 2 shows that this is the case for log(E) ?< -2, where the observed frequencies of E for the 70 DNA queries are similar to those for the 7082 protein queries.

There were very few observed E-values with log(E) = -2 and none with log(E) = -1 for the DNA queries. This is a consequence of the DNA query length. The probability for each supporting block in a hit includes a distance component, which is the ratio of a function of the observed range of distances between the supporting block and anchor block among sequences in the Blocks Databases and the query sequence length. For the DNA query, this ratio averages 3 x 126 / 315,357 = 1.2×10^{-3} (126 is the average value of the distance function in the Blocks Database in amino acid units). For log(E) to be -2, the distance function would have to be over 1000, which occurs rarely in the Blocks Database, and for log(E) to be -1 the distance function would have to be over 10,000, which never occurs.

So when applied to long DNA sequences the Blocks Searcher under-reports the less significant multiple block hits due to the high background levels, whereas it reports the more significant hits

with realistic expectant values, similar to those observed for protein queries. We can therefore use Table 1 to evaluate searches of the Blocks Database with long DNA queries.

3.2: Detection of homologs to yeast chromosome III ORFs

The longest available sequence to date is that of yeast chromosome III, on which 182 separate open reading frames (ORFs) have been identified [6]. Of these, 37 had been previously determined. For 30 other ORFs, significant similarities to known genes were detected based on FASTA scores of ≥200 when each was searched against sequence databanks. In a subsequent study another group used "a combination of low-stringency sequence database searches, various ways of assessing significance, multiple sequence alignment, pattern searches and incorporation of prior knowledge about protein and domain families" to search individual ORFs for additional relationships [7], [8]. This resulted in detection of 17 more relationships not reported in the first study.

For our test the entire 315,357 base pair (bp) sequence was used as query, with translation of all 6 reading frames. So that these results would be comparable to those from the earlier studies, a contemporary version of the Blocks Database (BLOCKS v. 4.1) was searched. The rank-ordered list of hits reported by BLOCKSORT is shown in Table 3a. The top 21 hits identified alignments that were previously known or first reported in the original study. Among these were hits involving multiple blocks for which expectant values are reported, as well as very high-scoring single block hits. Hit #22 was for the (cyclophilin-type) peptidyl-prolyl cis-trans isomerases, aligning with ORF R69w, an essential gene ($E=9x10^{-7}$ for 2 of 2 blocks). Truncation of R69w within the region that aligns with the second of two blocks suggests a possible sequencing error. Hit #25 was a spurious single block hit. Hit #26 was for the DNA polymerase X proteins, which aligned with segments from ORF R14c ($E=9x10^{-7}$ for 3 of 5 blocks). Hit #29 was for the prokaryotic carbohydrate kinases, which aligned with a segment of ORF R36w ($E=1x10^{-5}$ for 2 of 4 blocks). Hit #30 was for a single block representing fungal regulatory proteins with a Zn(2)-Cys(6) cluster, which aligned with a segment of ORF R106w. Like several members of this group such as ARG2, GAL4, and LEU3 from *S. cerevisiae*, R106w is 800-900 aa in length, with the block alignment falling within the first 50 residues. Hit #33 was for the sugar transport proteins, which aligned with ORF R98c ($E=7x10^{-4}$ for 2 of 4 blocks), and hit #42 was for the zinc-containing alcohol dehydrogenases, which aligned with ORF R102c ($E=7x10^{-4}$ for 2 of 4 blocks, A) The top 30 hits are shown as well as all multiple block hits among the top 316. Homologies not reported in studies by Bork and co-workers [7], [8] are indicated as new*. For hits 7 and 12, no E was reported because a higher scoring hit was mapped on the same strand. Examples in which Internal repeats were detected are indicated (†). Hits 21 and 34 were not detected in the search of BLOCKS v. 6.0. For hit 30, the undetected block was weak (Strength<1100). B) $G_{1-3}T$ telomeric repeat translated into Cys-, Gly-rich repeat is similar to an integrin repeat ($^{\#}$). Fig. 3 a). ORF R102c was one of the 7082 test queries in our previous study [18], ($E=1.4x10^{-6}$ for 3 of 4 blocks there). The failure to detect the third block in the present study is due to truncation of the results list to the top scoring 5000 alignments.

Five of the six hits (Table 3a) not reported in the original study were found in the second more detailed examination [7], [8]. However, the identification of R102c as a member of the zinc-containing alcohol dehydrogenases was not reported in either study. Among the top 316 hits only a single false positive multiple block hit was reported, the 61st best hit consisting of only 2 of 15 blocks in the best path ($E=7x10^{-3}$).

The 34 non-spurious hits listed in Table 3a represent about 40% of the 84 total homologies reported in previous studies of yeast chromosome III sequences; this is comparable to the fraction of SWISS-PROT sequences represented in the Blocks Database. Therefore, the detection of only a subset of known homologs in this search is accounted for by the fact that most proteins in sequence databases have no known homologs and so have not been placed into PROSITE groups.

```
A) Hit 42: Zinc alcohol dehydrogenases (Detection of distant relationship)

Block    Rank Frame Score Strength Location
BL00059A   78   -3   1302 2708     3720-3764
BL00059D  778   -3   1122 2214     3890-3947

1302=98.30th percentile of anchor block scores for shuffled queries
E=0.00074 for BL00059D in support of BL00059A
                        |-----  124 residues----|
   BL00059 AAAAAAAAA:.BBBBBB:.......CCCCCCCCCCCC:..........DDDDDDDDDDD
 YSCCHRIII AAAAAAAAA:::::::::::::::::::::::::DDDDDDDDDDD

BL00059A    <->A (1,35):3719
ADHX$HORSE 8     AAVAWEAGKPVSIEEVEVAPPKAHEVRIKIIATAVCHTDAYTLSG
                  ||  | || |  | |         | ||  | |   ||
YSCCHRIII  3720  KAVVIEdGKaVVkEgVPiPELeEGfVLIKtLAVAgnpTDwahIDy

BL00059D   A<->D (108,194):125
ADH$ARATH  198   AIFGLGAVGLGAAEGARIAGASRIIGVDFNSKRFDQAKEFGVTECVNPKDHDKPIQQV
                    |  |||      |        | |    |     || |        | |   |
YSCCHRIII  3890  LwgGAtAVGqSLIQlAnKLnGftkIIVvASrKhEKLlKEYGADqlfDyhDiDvveQIk

_____________________________________________________________________

B) Hit 23: Eukaryotic RNA-binding region (Repeated domains)

Block    Rank Frame Score Strength Location
BL00030A   50   -2   1530 1513     70920-70940
BL00030A  109   -2   1259 1513     71017-71037
BL00030A 1092   -2   1104 1513     71147-71167
BL00030B   80   -2   1292 1668     71054-71072
BL00030B  196   -2   1207 1668     70956-70974

1530=99.97th percentile of anchor block scores for shuffled queries
E=8.5e-05 for BL00030B in support of BL00030A

BL00030A    <->A (8,571):70919
SSB1$YEAST 37    TIFIGNVAHECTEDDLKQLFV
                  ||  |    ||  ||| ||
YSCCHRIII  70920 SIFVRNLtFDCTpEDLKELFG

BL00030B   A<->B (3,295):113
ROA1$HUMAN 51    RSRGFGFVTYATVEEVDAA
                  ||||| | | |  |   |
YSCCHRIII  71054 fSRGFGsViYpTEDEMIrA
```

Fig. 3. Block Searcher output for examples discussed in the text. For each alignment, the rank, frame, PATMAT score, strength and location are reported, where location refers to the aligned segment in the query. Alignments with scores that rank among the top 1264 in the search are mapped in panel A for the blocks in the hit. Minimum (:) and maximum (.) distances between blocks in the database are indicated. In the alignments, precise distances between detected blocks is shown as (min, max): for the database sequences followed by the distance in the query sequence.

We also used the same query to search the current Blocks Database, BLOCKS v. 6.0, which is one year newer and contains 23% more groups than BLOCKS v. 4.1 used in the above analysis (619 groups vs. 504 groups). This search classified an additional ten ORFs, missing two ORFs

detected previously (Table 3b), apparently because of changes to the blocks that occurred with the addition of new members to the group. The net increase of 8 ORFs detected (24%) is comparable to the increase in the number of families represented in BLOCKS after a one-year interval. While seven of the classifications represent findings listed by Oliver *et al.* [6], and one was missed by them but was reported in the subsequent studies by Bork and co-workers [7], [8], two others were not detected by either group. One is a new member of the beta-transducin family (R57c), and the other is an EGF-related protein. In both cases, identification was confirmed by the detection of an adjacent repeated domain (see below). ORF R57c was one of the 7082 queries used in our earlier study [18], detecting the two beta-transducin family blocks with $E=3x10^{-5}$, similar to the level of detection in the present study ($1x10^{-4}$, Table 3b).

One interesting false positive detected in this search is the translated sequence of telomeric $G_{1-3}T$ DNA repeats aligned with blocks representing an integrin beta chain repeat unit. This is accounted for by the fact that the integrin blocks consist largely of cysteine and glycine residues which can be encoded by TGT and GGN respectively. Like a similar example reported previously for a region of unusual base composition [17], the basis for this block-specific artifact is readily identified by examination of the BLOCKSORT output. Alternatively, such artifacts might be reduced or eliminated by using sequence filters to remove regions of low informational complexity from sequence queries [20], [21].

In five cases, we detected characteristic repeated motifs, seen as multiple alignments of a single block within the same encoded protein. An example is Hit #23, the eukaryotic RNA-binding domain typically found 2-3 times in proteins of this family (Fig. 3b). In the case of the EGF domain reported in the search of BLOCKS 6.0 for ORF R11c, the detection of a very high-scoring repeat (at the 98th percentile of anchor block scores) beginning 38 aa upstream is strong confirmation of what otherwise would have been a "twilight zone" hit (at the 99.1 percentile for a single block). ORF R11c is unusual in that there are two clear family relationships within the same predicted protein, one belonging to the EGF family described here, and the other belonging to the ATP-dependent transporters, detected previously [6]. Since the EGF homology is within the first 200 aa, while the ATP-dependent homology is within the last 700 aa of a 1049 aa protein, it is possible that a fusion has occurred. However, whether this apparent fusion resulted from a real evolutionary event or a sequencing error remains to be determined.

Table 3. Highest scoring block hits in S. cerevisiae chromosome III

Hit #	Family	Percentile	ORF	Frame	#Blocks	E	Comment	Zone
A. Versus BLOCKS v. 4.1 (Feb. 1992)								
1	Phosphoglycerate kinases	100	L12w	2	8/8	$6x10^{-36}$	known	<1/7000
2	Citrate synthases	100	R5c	-1	6/6	$4x10^{-24}$	known	<1/7000
3	Hexokinases	100	R40w	2	5/5	$3x10^{-21}$	known	<1/7000
4	Isocitrate dehydrogenases	100	L18w	2	5/5	$5x10^{-20}$	known	<1/7000
5	Ribosomal protein S11	100	R31c	-2	2/2	$3x10^{-6}$	known	<1/7000
6	Protein kinases	100	L24w	3	2/2	$2x10^{-4}$	reported	<1/7000
7	Protein kinases	100	L8w	3	2/2		reported	<1/7000
8	Thioredoxins	100	L43c	-1	2/2†	$2x10^{-4}$	reported	<1/7000
9	Protein kinases	100	R73c	-2	2/2	$1x10^{-4}$	reported	<1/7000
10	ATP-binding transporters	100	R11c	-2	2/2	$3x10^{-4}$	reported	<1/7000
11	Thioredoxins	100	R83w	1	2/2	$3x10^{-5}$	reported	<1/7000
12	Protein kinases	100	R91w	1	2/2		reported	<1/7000
13	Amino acid permeases	100	L25c	-1	2/2	$9x10^{-6}$	reported	<1/7000
14	Glutaredoxins	100	R36w	-2	2/2	$6x10^{-6}$	reported	<1/7000
15	Homeobox	100	R97w	3	1/1		known	<1/7000
16	Ser/Thr dehydratases	100	L64c	-1	2/2	$4x10^{-5}$	reported	<1/7000
17	Serine proteases	100	R45c	-2	4/4	$1x10^{-15}$	reported	<1/7000
18	Homeobox	100	R96c	-1	1/1		known	<1/7000
19	Homeobox	100	L67c	-2	1/1		known	<1/7000
20	Homeobox	100	R39c	-3	1/1		known	<1/7000
21	Endoplasmic ret. target	100	L43c	-1	3/3	$6x10^{-7}$	known	<1/7000
22	Pept.-prolyl isomerases	99.97	R69w	3	2/2	$9x10^{-7}$	new	<1/7000
23	RNA-binding proteins	99.97	L11c	-2	2/2†	$9x10^{-5}$	reported	<1/7000
24	Zinc-containing ADHs	99.77	R105w	1	3/4	$2x10^{-8}$	reported	<1/7000
25	Engrailed-type homeobox	99.74		2	1/7		spurious	>1/100
26	DNA polymerase X proteins	99.74	R14c	-1	3/5	$9x10^{-7}$	new	<1/7000
27	Class I metallothionines	99.70		-2	1/1		spurious	>1/100
28	Receptor tyrosine kinases	99.70	R73c	-2	1/5		reported	>1/100
29	pfkB carbohydrate kinases	99.57	R36w	1	2/4	$1x10^{-5}$	new	<1/7000
30	Fungal Zn-Cys cluster	99.39	R106w	2	1/2		new	>1/100
33	Sugar transport proteins	99.19	R98c	-1	2/4	$7x10^{-4}$	new	<1/7000
34	C2 domain	99.18	R91w	1	2/4	$6x10^{-5}$	reported	<1/7000
37	Phorbol ester binding	98.80	R91w	1	2/3	$8x10^{-5}$	reported	<1/7000
42	Zinc-containing ADHs	98.30	R102c	-3	2/4	$7x10^{-4}$	new*	<1/7000
49	Receptor tyrosine kinases	97.49	L24w	3	2/5	$5x10^{-4}$	reported	<1/7000
61	Euk. RNA polymerase II	96.40		3,1	2/1	$7x10^{-3}$	spurious	>1/1000
115	Phorbol ester binding	90.60	R73c	-2	2/3	$4x10^{-4}$	reported	<1/7000
B. Additional top hits versus BLOCKS v. 6.0 (Feb. 1993)								
	Fork head domain	100	R65w	2	3/3	$1x10^{-11}$	reported	<1/7000
	Histidinol dehydrogenase	100	L30c	-3	8/8	$1x10^{-35}$	known	<1/7000
	DNA mismatch repair	100	R92c	-3	2/2	$3x10^{-5}$	reported	<1/7000
	Stress-induced proteins	100	R104w	-1	4/4	$6x10^{-13}$	new	<1/7000
	G protein	100	R38c	-3	2/2	$2x10^{-3}$	reported	<1/7000
	Beta transducin	100	R57c	-3	2/2†	$3x10^{-5}$	reported	<1/7000
	C3-C4 Zinc finger	100	R66w	2	1/1†		known	<1/7000
	Beta transducin	99.95	L39w	1	2/2	$1x10^{-4}$	new*	<1/7000
	Integrins beta chain	99.94		-2	2/8	$4x10^{-4}$	telomere#	<1/7000
	Neutral metallopeptidase	99.65	L57w	2	1/1		known	>1/100
	EGF-like	99.10	R11c	-2	1/2†		new*	>1/100

4: Discussion

In previous work, we described an automated approach to protein family classification that involves using a protein query to search a database of protein blocks for both local and global similarities [18]. The database includes families that are usually represented by more than one block; therefore detection of multiple blocks can be used to compute a global "expectant value". This value can be combined with an independent local "anchor score" to arrive at an overall level of confidence. Here, we have applied this method to analyze a very large dataset encoding numerous proteins, choosing the largest sequence available, yeast chromosome III. While we recognize that typical genomic sequences from higher eukaryotes will be more challenging because coding regions are separated by introns, we think that our general approach is applicable to that situation. In particular, the short conserved regions represented by blocks should be appropriate for detection of homology within exons, and the examination of all 3 frames on a strand should allow for multiple block alignments within different exons. In such applications, it might be worthwhile to use a more flexible function for the allowable distance between blocks so that gaps caused by introns can be bridged.

An advantage to our automated approach is the ready detection of repeats by examination of BLOCKSORT output (see Fig. 3b). Two of the three new homologies detected here but not in previous studies were of this type. This suggests that manual methods in general use are relatively insensitive to repeats.

The Blocks Searcher is extremely sensitive to distant relationships. For yeast chromosome III, the search of BLOCKS v. 4.1 involved more than 10^9 block alignments, necessitating a high threshhold for detection. Nevertheless, this single search detected a relationship to an ORF (R102c) that was missed by two groups examining the results of searches using 182 queries. This example also illustrates that it is not necessary to identify ORFs in order to do an adequate search. As a result, a hit involving multiple blocks in different reading frames can be detected and statistically evaluated, which is an important feature for family classification involving raw and erroneous DNA sequence [2].

The Blocks Searcher is also selective when applied to large DNA sequences. This is illustrated by the observation that there was only a single spurious hit involving multiple blocks among the top 316 hits reported using yeast chromosome III as query. In one sense, this low background resulted from the high threshhold necessitated by the sheer size of the search. While it is possible that true positives were missed because of the high threshhold, it is comforting that we were able to detect the same fraction of homologs detected by others (40%) as the fraction of SWISS-PROT sequences present in the PROSITE catalog from which the Blocks Database was derived. When the same search was carried out a year later with a larger Blocks Database, there was a corresponding increase in the number of homologs detected. As the Blocks Database expands slowly in the number of groups, our approach becomes more powerful; in contrast, the sequence databases expand much more rapidly, and detection of homology becomes more difficult due to increased background.

5: Conclusion

Our approach to the analysis of large-scale sequence data provides an automated method for the detection and evaluation of family relationships. In contrast to manual approaches in which individual open reading frames are first identified, then searched and evaluated individually, we carry out a single search of the entire sequence. Yeast chromosome III provided an excellent test of our approach, because it has been the subject of intensive analysis by groups of sequence analysis experts [6], [7], [8]. Our ability to detect clear relationships not reported in any of those studies argues that an automated approach can be very powerful. No special expertise is necessary to use this system; about 1500 people have used it as an electronic mail server.

6: Acknowledgments

This work was supported by a grant from NIH (RO1 GM29009). Part of this work was performed in the 1992 Recognizing Genes Workshop at the Aspen Center for Physics.

7: References

[1] R. F. Smith and T. F. Smith "Automatic generation of primary sequence patterns from sets of related protein sequences," Proceedings of the National Academy of Sciences of the United States of America, vol. 87, pp. 118-122, 1990.

[2] S. Henikoff and J. G. Henikoff "Automated assembly of protein blocks for database searching," Nucleic Acids Research, vol. 19, pp. 6565-6572, 1991.

[3] A. Ogiwara, I. Uchiyama, Y. Seto and M. Kanehisa "Construction of a dictionary of sequence motifs that characterize groups of related proteins," Protein Engineering, vol. 5, pp. 479-488, 1992.

[4] N. Harris, L. Hunter and D. States, "Megaclassification: discovering motifs in massive datastreams," in Proceedings of the National Conference on Artificial Intelligence Menlo Park: AAAI Press, 1992, pp. 224-232.
[5] R. P. Sheridan and R. Venkataraghavan "A systematic search for protein signature sequences," Proteins: Structure, Function and Genetics, vol. 14, pp. 16-28, 1992.
[6] S. G. Oliver, Q. J. M. van der Aart, M. L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar and J. P. G. Ballesta "The complete DNA sequence of yeast chromosome III," Nature, vol. 357, pp. 38-46, 1992.
[7] P. Bork, C. Ouzounis, C. Sander, M. Scharf, R. Schneider and E. Sonnhammer "What's in a genome?," Nature, vol. 358, pp. 287, 1992.

[8] --- "Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III," Protein Science, vol. 1, pp. 1677-1690, 1992.
[9] T. F. Smith and M. S. Waterman "Identification of common molecular subsequences," Journal of Molecular Biology, vol. 147, pp. 195-197, 1981.
[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman "Basic local alignment search tool," Journal of Molecular Biology, vol. 215, pp. 403-410, 1990.
[11] W. R. Pearson "Rapid and sensitive sequence comparison with FASTP and FASTA," Methods in Enzymology, vol. 183, pp. 63-98, 1990.
[12] M. Gribskov, R. Luthy and D. Eisenberg "Profile analysis," Methods in Enzymology, vol. 183, pp. 146-159, 1990.
[13] A. Bairoch "PROSITE: A dictionary of sites and patterns in proteins," Nucleic Acids Research, vol. 20, pp. 2013-2018, 1992.
[14] A. Bairoch and B. Boeckmann "The SWISS-PROT protein sequence data bank," Nucleic Acids Research, vol. 20, pp. 2019-2022, 1992.
[15] H. O. Smith, T. M. Annau and S. Chandrasegaran "Finding sequence motifs in groups of functionally related proteins," Proceedings of the National Academy of Sciences of the United States of America, vol. 87, pp. 826-830, 1990.
[16] S. Henikoff, J. C. Wallace and J. P. Brown "Finding protein similarities with nucleotide sequence databases," Methods in Enzymology, vol. 183, pp. 111-132, 1990.
[17] J. C. Wallace and S. Henikoff "PATMAT: a searching and extraction program for sequence, pattern, and block queries and databases," Computer Applications in the Biosciences, vol. 8, pp. 249-254, 1992.
[18] S. Henikoff and J. G. Henikoff "Protein family classification based on searching a database of blocks," Genomics, 1993, In press.
[19] S. Henikoff, J. G. Henikoff, S. Agus and J. C. Wallace, "Searching for homologies to protein blocks by electronic mail," in Automated DNA sequencing and analysis techniques (J. C. Venter, Ed.). London: Academic Press, 1993,.
[20] J. C. Wootton and S. Federhen "Statistics of local complexity in amino acid sequences and sequence databases," Computers and Chemistry, 1993, vol. 17, pp. 149-163.
[21] J. M. Claverie and D. J. States "Information enhancement methods for large scale sequence analysis," Computers and Chemistry, 1993, vol. 17, pp. 191-201.